MEDICAL RESEARCH LABORATORY REPORT NO. 57

MEDICAL RESEARCH DEPARTMENT

U. S. Submarine Base
New London

Vol. IV No. 1

FUNCTIONS OF LOUDNESS DISCRIMINATION IN SUBMARINE SONAR OPERATION

(Interval Report)
Bureau of Medicine and Surgery
Research Division - Project X-53
SELECTION OF SOUND OPERATORS

Prepared by

Ensign J. Donald Harris, H(S) USNR

5 April 1945

APPROVED: Captain C. W. Shilling, (MC) USN, MO-in-C

FUNCTIONS OF LOUDNESS DISCRIMINATION

IN

SUBMARINE SONAR OPERATIONS

(Interval Report)
Bureau of Medicine and Surgery
Research Division - Project X-53
SELECTION OF SOUND OPERATORS

Prepared by

Ensign J. Donald Harris, H(S) USNR

5 April 1945

APPROVED: Captain C. W. Shilling, (MC), USN, MO-in-C.

FUNCTIONS OF LOUDNESS DISCRIMINATION

IN SUBMARINE SONAR OPERATIONS.

TABLE OF CONTENTS

TABLES AND ILLUSTRATIONS

SUMMARY

1. A resume is given of the research on the relationship of pure-tone loudness discrimination to the submarine sonar performance.

2. The experience of this laboratory since July 1944 with Auditory Test No. 7 of the Harvard Psycho-Acoustic Laboratory, "Loudness Discrimination for Bands of Noise" is presented.

3. This test, which requires a subject to make 110 judgments as to whether a complex tone (500-2000 c.p.s.) becomes louder or softer in intensity, is satisfactorily reliable when administered as a group test with loudspeaker (odd-even r = +.88).

4. Performance on the test is independent of overall intensity level over a rather wide range.

5. The reliability coefficient of a short form of the test (first 80 items only) is .82, a drop of only .06 from that on the full test.

6. An item analysis is presented showing which items are effective in screening out poor individuals.

7. The psychological function relating performance to level of difficulty of items, is drawn for three representative groups.

8. The negative time-error for an average group is .27 decibels. This means that if the Standard Stimulus is presented, and followed by a Variable Stimulus .27 db <u>softer</u> in intensity, average subjects will nevertheless judge the sounds as of equal loudness. For poor subjects, the magnitude of the negative time-error is .6 db.

9. The physiological nature of the negative time-error is conjectured.

10. A statistical device is presented whereby a subject's differential loudness threshold can be estimated in db, merely from his raw score on the loudness test.

11. The relation of the test to sonar performance is investigated in preliminary experiments. No correlation exists between the test and final sound school grades; but when correlated against specific auditory sonar performances to which loudness discrimination may reasonably be presumed to contribute, correlations of the order .21 - .51 were obtained. In addition, significant differences in performance were found between those who do poorly and those who do average or better on the test.

12. A fuller report on these validation experiments is in progress.

******

FUNCTIONS OF LOUDNESS DISCRIMINATION IN SUBMARINE SONAR OPERATION

INTRODUCTION:

The first use of a test of loudness discrimination for the preselection of sound listeners was in August, 1941, when this activity under the direction of Doctor C. W. Shilling and Lt. I. A. Everley began routinely presenting the Seashore Series "B" Intensity Records to groups of submariners. In January, 1942, the easier Series "A", New Edition Seashore Intensity Record was substituted and routine testing continued.

With the additional assistance of Mrs. Jessie Kohl and Dr. W. D. Neff, this activity by March, 1942, had given the New Edition Seashore Series "B" Intensity Record to 204 subjects, and 337 subjects had been given the New Edition Series "A" record.

In January 1942, the Committee on Selection and Training of Sound Operators, Section C-4 of NDRC, presented the Old Edition Seashore Intensity Record to 65 student operators at the San Diego surface fleet sound school.

In April, 1942, the Series "A", New Edition record was correlated against the performance of 15 separate groups in three separate New London sound schools. The correlations ranged from -.12 to +.63. Inasmuch as the groups each contained only from 8 to 20 men, these results were inconclusive. In June 1942, we reported the Series "A" New Edition record and overall sound school performance to show biserial correlation coefficients between r = +.14 and +.20, for 210 subjects.

The uniformly insignificant validities reported above, as well as the rather low reliability commonly found, prompted all activities selecting sonar personnel to discontinue experimental work with the Seashore record.

It would seem odd, however, if so fundamental a factor as loudness discrimination were totally unrelated to soundman performance. And when it is realized that none of the validation material up to the time of discontinuing the Seashore record had included any job analysis of sonar performance (so far as the effort goes to isolate and test an aspect or aspects of that performance which might depend at least in part on loudness discrimination), it becomes plain that the relation of loudness to sonar performance had been by no means fully explored.

Although it was perhaps established that the Seashore records were of little value for the purpose, it remained possible that some other test of loudness discrimination would serve. Should such a test appear, this activity hoped to explore the possibilities immediately.

In July 1944, the Psycho-Acoustic Laboratory of Harvard University forwarded their Auditory Test No. 7, Loudness Discrimination for Bands of Noise. The determination of its characteristics in a military situation and its applicability to submarine sonar selection was at once begun. The rest of this paper describes the experience to date of our laboratory with Auditory Test No. 7.

TEST AND EQUIPMENT:

Test No. 7 consists of four 12-inch record sides, containing 110 items. Each item is a complex noise presented for 2 seconds, then increasing or decreasing in overall intensity for a similar time interval. Subjects are

required to judge whether the second half of the noise is "Louder" or "Softer" than the first half. The spectrum of noise used is flat over the range 500-2000 cycles per second, and on either side of this range the spectrum falls off at the rate of 17 decibels per octave. The change in <u>intensity</u> which occurs in the middle of each item is not accompanied by changes in the <u>frequency</u> composition of the noise (i.e., the item does not raise or lower in pitch).

The test combines the advantages of a reliable group test with the possibility that the test of loudness discrimination which depends upon a complex rather than a pure tone, might well be related to sonar performance.

All of the data reported in this paper were collected in a soundproof room 18' x 13' x 9' high, lined with a highly absorbent limestone tile. The noise attenuation from outside is in excess of 90 db.

The playback used was a Presto Model "L", coupled to a Jensen 12" dynamic speaker. The characteristics of this system easily meet the minimum requirements for administration of the test as recommended by the Psycho-Acoustic Laboratory.

The intensity of the test items was set with the aid of a General Radio Sound Level Meter, Type 759.

ADMINISTRATION:

After supplying subjects with pencils and answer blanks and assuring them it will be to their benefit to do their best, the record with its self-contained instructions is presented, without further comment. The output to the loudspeaker was adjusted to produce 85 decibels

as the average sound level meter reading at 10 representative positions in the test room.

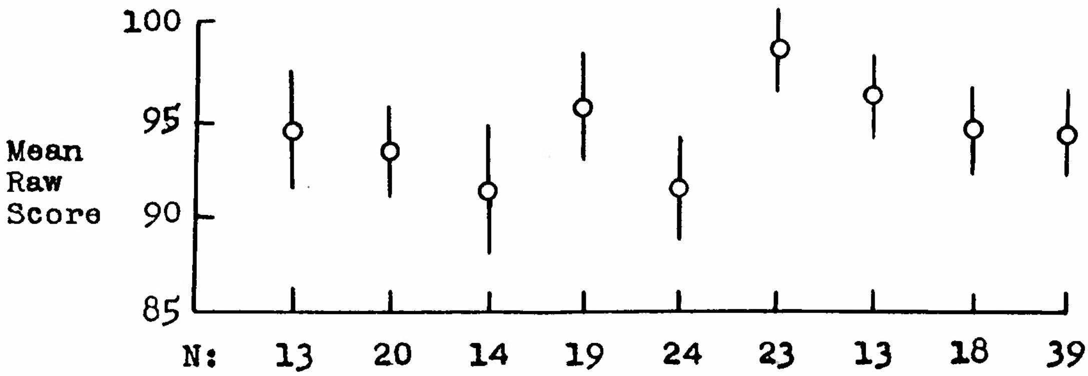
Figure 1 shows the variations from day to day in the mean score of consecutive routine groups. A certain stability has been achieved, on the basis of which a frequency distribution was constructed, its mean, standard deviation, and standard error computed.

Figure 2 presents the frequency distribution for 1,060 subjects given our routine administration.

From this data it is possible to assign a T-score to every possible raw score according to the formula $T = 50 + \frac{x}{\sigma} 10$, where x = deviation of raw score from the mean. Thus, a T-score of 60 represents a raw score better than average by one standard deviation, a T-score of 40 represents a raw score worse than average by one standard deviation, and so on.

RELIABILITY:

The reliability coefficient for our groups, relating a man's score for all odd items against that for all even items, is .89, corrected for length (N = 281). (For the purposes of this laboratory it is the characteristics of a first administration that must be known, consequently odd-even reliability coefficients rather than test-retest are of first importance.) Since our subjects are preselected for intelligence (Mean IQ = 110, with cases only rarely below 90), it is almost certain that the reliability of this test, as we administer it, would be even higher with a wider range of intelligence.



Note: Vertical lines = ± 2 Standard Errors

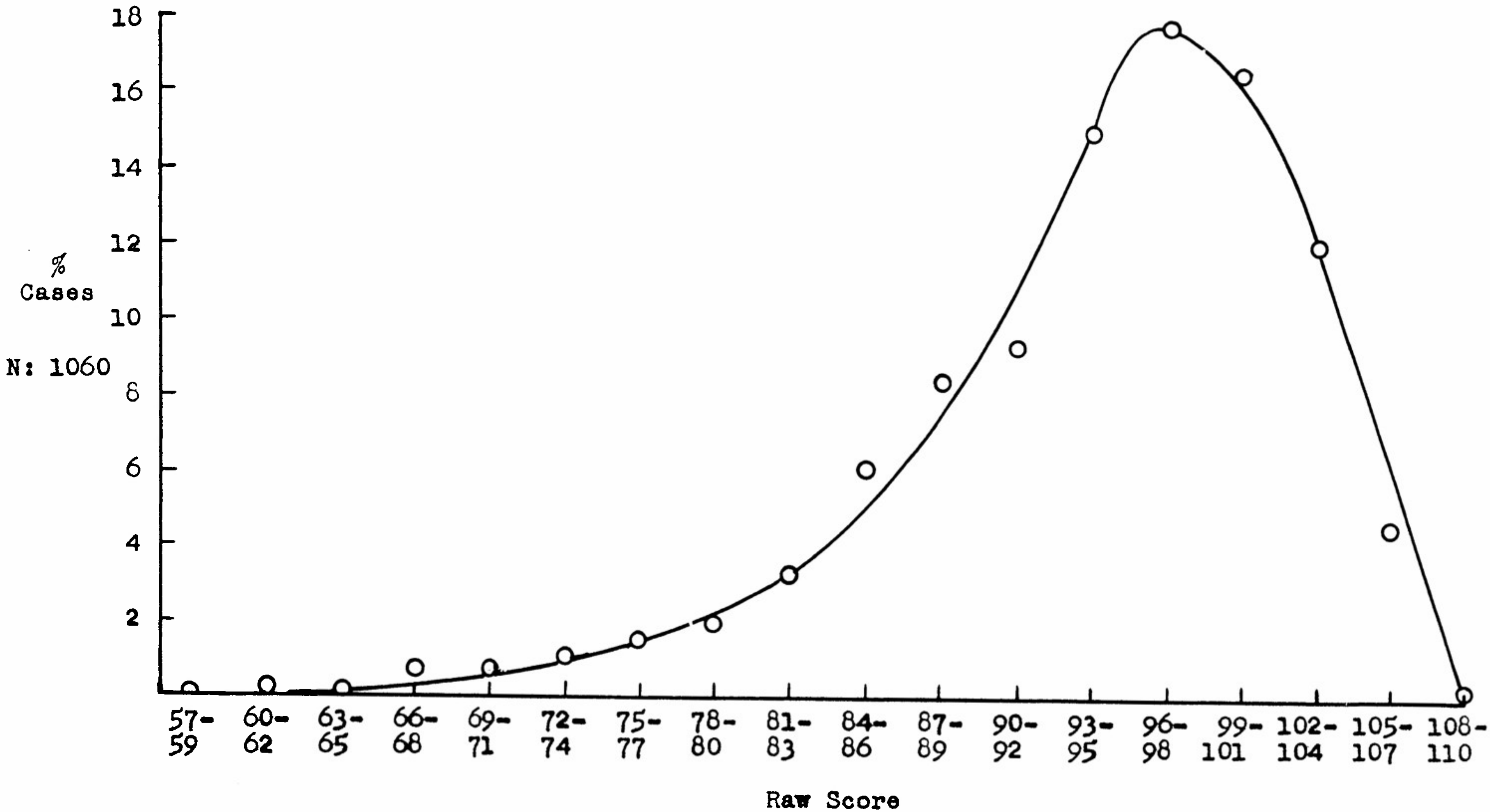
Mean Scores for Consecutive Daily Groups

Fig. 1

Fig. 2

Frequency Distribution of Loudness Test

Percent of Cases at each Score Interval



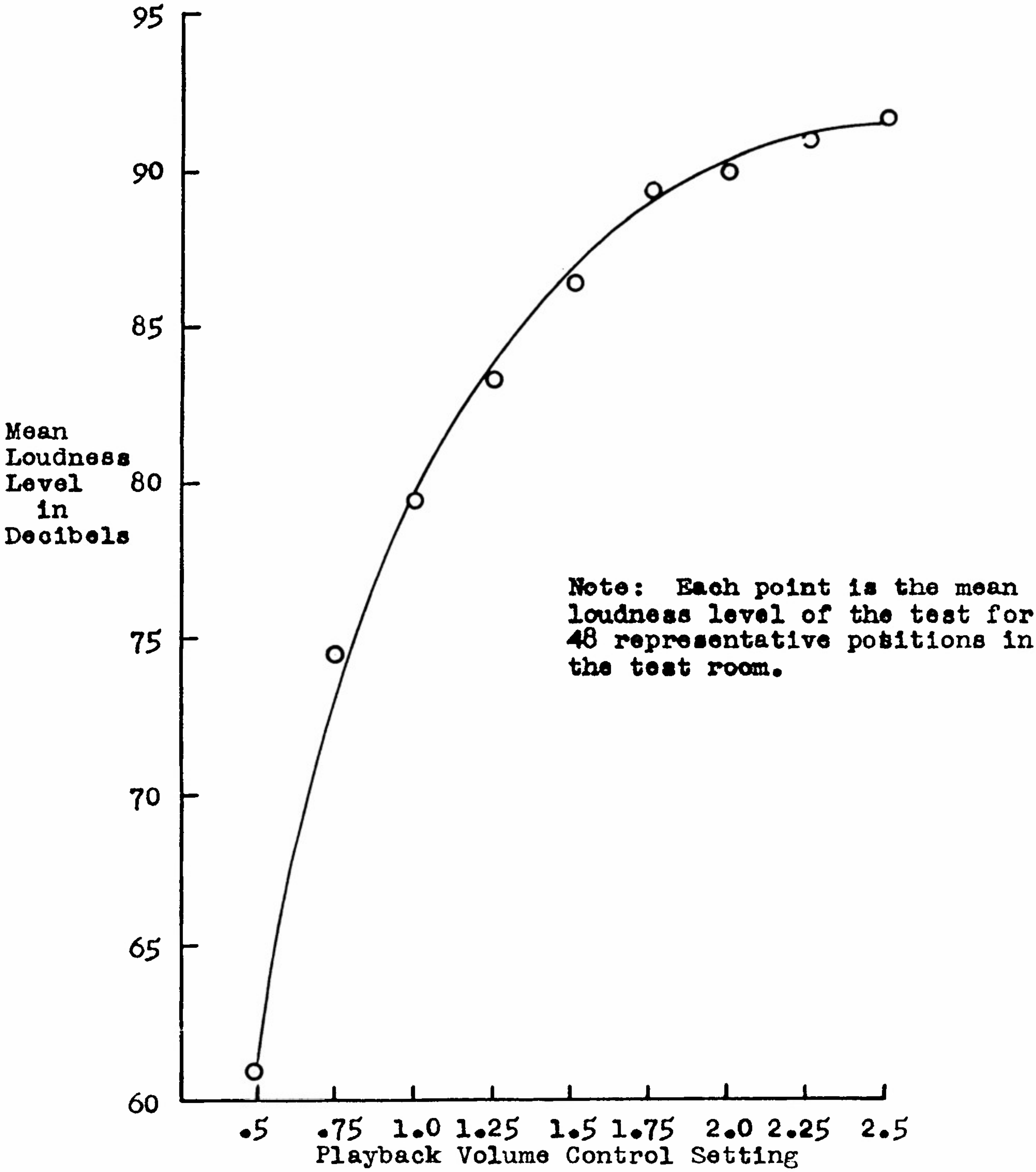
For headphone administration the highest correlation reported for this test by the Psycho-Acoustic Laboratory is .85 for one group of 53 subjects. It may be that our slightly higher figure is due to the administration by loudspeaker, but it is more likely due to the probability that in our somewhat larger group a greater range of ability was present. There seems to be no very important difference between headphone and loudspeaker administration, nor from group to group, so far as reliability is concerned.

The question arises, what is the effect of seating position of the subject with respect to the loudspeaker. It seems that the position is largely immaterial: within the range of loudness variation from seat to seat, changes in loudness level of test administration are not followed by important changes in test performance. The experimental argument follows:

The range of loudness variation from seat to seat is given in Figure 3; a scale drawing of the seating arrangement is shown with the average loudness level for each seat entered on the drawing. The level entered for each seat is an average of 3 separate readings on the sound level meter, at each of nine different volume control settings on the playback.

A range of 11.5 db can be seen in Figure 3, from seat #8 to #5. The lowest average was at seat #8; from the figures at the other seats can be computed the db by which the averages were louder at those seats. In order to determine whether seating arrangement is a definite factor, it is then only necessary to discover whether, through a range of approximately 11.5, test performance is affected.

The relation between the volume control settings on the playback, and the average loudness level of the test in the room, was first ascertained. Figure 4 describes the relationship.



Mean Loudness Levels Produced by
Several Playback Volume Control Settings

Fig. 4

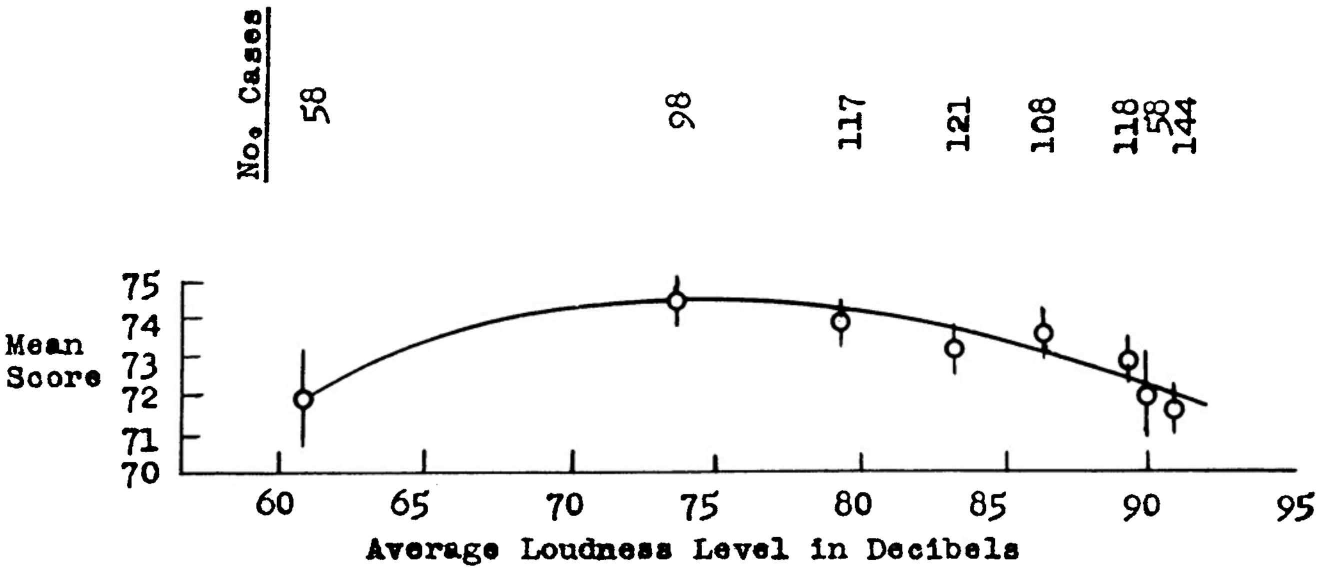
On the baseline are to be found the volume control settings on an arbitrary scale, while on the ordinate is found the average loudness level for 48 representative positions in the room.

The average test score when the test is administered at these average levels, is shown in Figure 5, (For this purpose a shortened form of the test, consisting of the first three of the four record sides, Items 1-80, was used. The characteristics of a shortened form are described in a later section). The vertical lines drawn through the plots represent ± 2 standard errors. Although a slight tendency exists for mean score to be best at 75-80 db and to fall off with the lower and higher levels, it is clear that loudness variations within our room are of relatively minor importance as a factor in performance. It is probably true, therefore, that loudspeaker administration will be satisfactory and will contribute to the efficiency of other activities wishing to make use of this test.

SHORT-FORM ADMINISTRATION:

In the hope of saving testing time, we interested ourselves in determining whether the 110 items could be reduced to some minimum and used in a short form of the test. This reduction seemed feasible especially as the more difficult items are not all concentrated at the end; Items 37-50, for example, contain samples as difficult as any in the whole test.

To produce a short form, the last record side, containing Items 81-110, was eliminated. The test thus administered contains a representation of items at all levels of difficulty. The only question would seem to be, whether the shortening has a very deleterious effect on reliability.



Note: Vertical line = ± 2 Standard Errors

Relation between Average Score and Loudness
Level at which Test is Run

Fig. 5

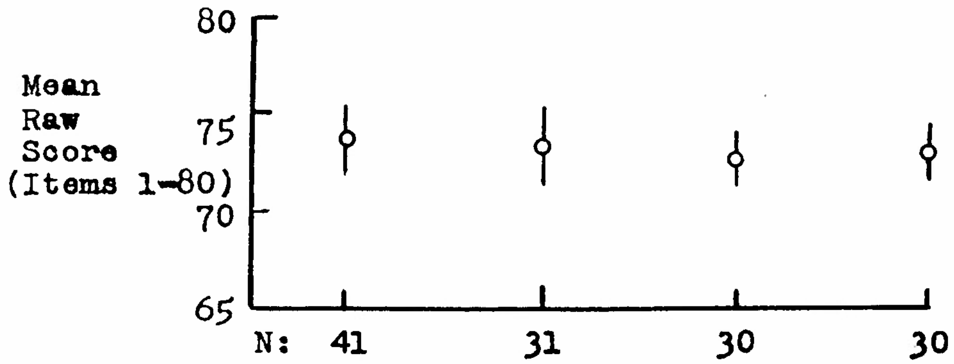
Reduction of 80 items does indeed lower reliability, but not perhaps too greatly. With 277 subjects, the odd-even reliability coefficient of a first administration was .82, a drop of .06 from that on the full test.

Figure 6 plainly shows that the short-form test produces similar results when given to similar groups. The day by day means are indistinguishable.

That the results of a short-form administration have the same general significance as the full test is shown by comparing a subject's score on the first 80 items with his score for the 110 items. The comparison yields a correlation of .84; in other words, the relation between the two scores is of about the same order as that for test-retest with the full test. We reason that for allpractical purposes the two forms are interchangeable.

ITEM ANALYSIS:

Since our primary purpose in using the test is in the elimination of those individuals who are poor in loudness discrimination, we were interested in determining which items were actually doing the eliminating. Accordingly an item analysis was performed, comparing the percentage of correct judgments (for each item separately) of two groups, one consisting of those subjects who scored about average ($\pm$ .5 S.D. above the mean), the other consisting of those who scored low (1 S.D. or more below the mean). Table I shows this analysis by items in terms of the difference in percentage correct between groups, the standard error of the difference, and the Critical Ratio of the difference. The statistical convention is that for a difference to be considered reliable, its C.R. must be at least 3.0 or better.



Note: Vertical lines = ± 2 Standard Errors

Mean Scores for Consecutive Daily Groups

Fig. 6

TABLE I

Auditory Test No. 7

Loudness Discrimination

for Bands of Noise
- Item Analysis -

Entry: Difference in percentage correct between a "Low" group and an "Average" group; standard error and critical ratio of that difference.

"Softer" Items				"Louder Items			
Db. Diff. = 3.1				Db. Diff. = 2.5			
Item No.	Diff.	S.E.	C.R.	Item No.	Diff.	S.E.	C.R.
1	4.1	2.48	1.7	2	4.7	2.17	2.2
4	10.8	2.48	4.3	3	18.9	3.27	5.8
Db. Diff. = 2.15				Db. Diff. = 1.55			
6	10.1	2.81	3.6	5	10.1	2.81	3.6
7	21.6	4.39	4.9	9	2.0	1.81	1.0
8	24.3	4.56	5.3	10	17.7	4.56	3.9
Db. Diff. = 1.55				Db. Diff. = 1.05			
11	13.5	2.81	4.8	13	15.5	3.24	4.8
12	16.3	4.86	3.4	15	8.1	3.51	2.3
14	13.5	3.88	3.5	16	7.5	4.56	1.6
22	26.5	6.43	4.1	21	23.7	8.86	2.7
23	28.8	8.32	3.5	24	17.6	3.70	4.8
32	18.3	5.67	3.2	31	19.7	5.97	3.3
52	22.3	4.24	5.2	51	20.9	6.35	3.3
53	35.9	6.51	5.3				
54	20.5	7.39	2.8				

TABLE I

"Softer" Items				"Louder" Items			
Db. Diff. = 1.05				Db. Diff. = .6			
Item No.	Diff.	S.E.	C.R.	Item No.	Diff.	S.E.	C.R.
17	16.2	3.05	5.3	18	27.0	6.38	4.2
19	21.7	5.67	3.8	20	6.9	3.88	1.0
26	42.1	8.63	4.9	25	17.7	5.79	3.1
27	21.9	7.53	2.9	28	25.8	4.96	5.2
29	19.9	4.44	4.5	30	20.5	7.09	2.9
33	24.8	8.04	3.1	34	15.6	5.79	2.7
35	33.9	7.22	4.7	55	21.7	5.10	4.3
36	12.6	7.74	1.6	56	27.2	7.09	3.8
57	26.3	6.43	4.1	58	12.3	6.08	2.0
59	26.4	5.95	4.4	60	13.6	6.75	2.0
62	10.2	5.32	1.9	61	4.3	7.10	6.1
71	4.8	5.10	0.9	72	15.7	6.89	2.3
Db. Diff. = .6				Db. Diff. = .1			
48	27.5	4.76	5.8	47	13.9	8.24	1.7
49	8.7	8.61	1.0	50	8.5	8.00	1.1
86	23.5	8.58	2.7	85	33.9	7.39	4.6
89	16.6	8.26	2.0	87	25.1	6.60	3.8
90	14.6	8.33	1.8	88	21.3	4.64	4.6
93	9.4	8.64	1.1	94	10.5	8.24	1.3
95	3.2	8.58	0.9	96	15.2	8.18	1.9
98	32.8	8.14	4.0	97	18.9	8.66	2.2
99	16.1	8.59	1.9	100	11.2	8.33	1.3
103	0.8	7.33	0.01	106	21.9	8.00	2.7
104	25.2	7.55	3.3	108	5.8	8.29	7.0
105	19.9	4.76	4.5	110	16.7	8.40	2.0
107	32.4	7.71	4.2				
109	17.9	4.64	3.7				

TABLE I

TABLE I (cont'd)

"Softer" Items Db. Diff. = .9				"Louder" Items Db. Diff. = .4			
Item No.	Diff.	S.E.	C.R.	Item No.	Diff.	S.E.	C.R.
39	23.4	8.10	2.9	37	22.5	7.22	3.1
41	28.0	8.18	3.8	38	5.4	8.00	0.7
43	27.4	7.84	3.5	40	27.2	7.09	3.9
46	23.8	6.74	3.5	42	21.8	6.97	3.1
63	7.7	7.12	1.1	44	9.2	4.79	1.9
65	23.1	6.43	3.6	45	7.8	4.63	1.7
67	27.8	6.67	4.2	64	8.2	5.67	1.4
68	17.8	7.61	2.3	66	19.8	7.51	2.6
73	19.0	5.57	3.4	69	12.4	7.21	1.7
76	19.4	6.35	3.1	70	8.4	7.79	1.1
77	27.4	7.88	3.5	74	22.6	7.96	2.8
79	8.4	7.51	1.1	75	28.8	8.17	3.5
83	24.0	8.48	2.9	78	19.8	6.81	2.9
84	10.5	8.24	1.2	80	28.4	7.28	3.9
92	26.1	7.09	3.7	81	12.3	5.57	2.2
101	11.6	6.36	1.9	82	20.3	7.09	2.9
				91	25.9	7.50	3.5
				102	1.4	6.35	0.2

Forty-eight items meet the criterion C.R. = 3, while thirteen more fall in the generally significant range 2.5 - 2.9. It is clear that the test contains sufficient items which are effective in screening out poor individuals.

The distribution of these critical items through the test is of some interest. Table I is constructed in such a way as to indicate the effect of increasing the difficulty of items. The number of critical items increases as the level of difficulty increases, up to a certain point. But it is seen that by no means all of the discriminative items are placed in the latter sections,--many are placed quite early. From the point of view whether a later section may be eliminated in the interest of efficiency, it will be noted that only 11 of the 30 items from No. 81-110 are clearly discriminative.

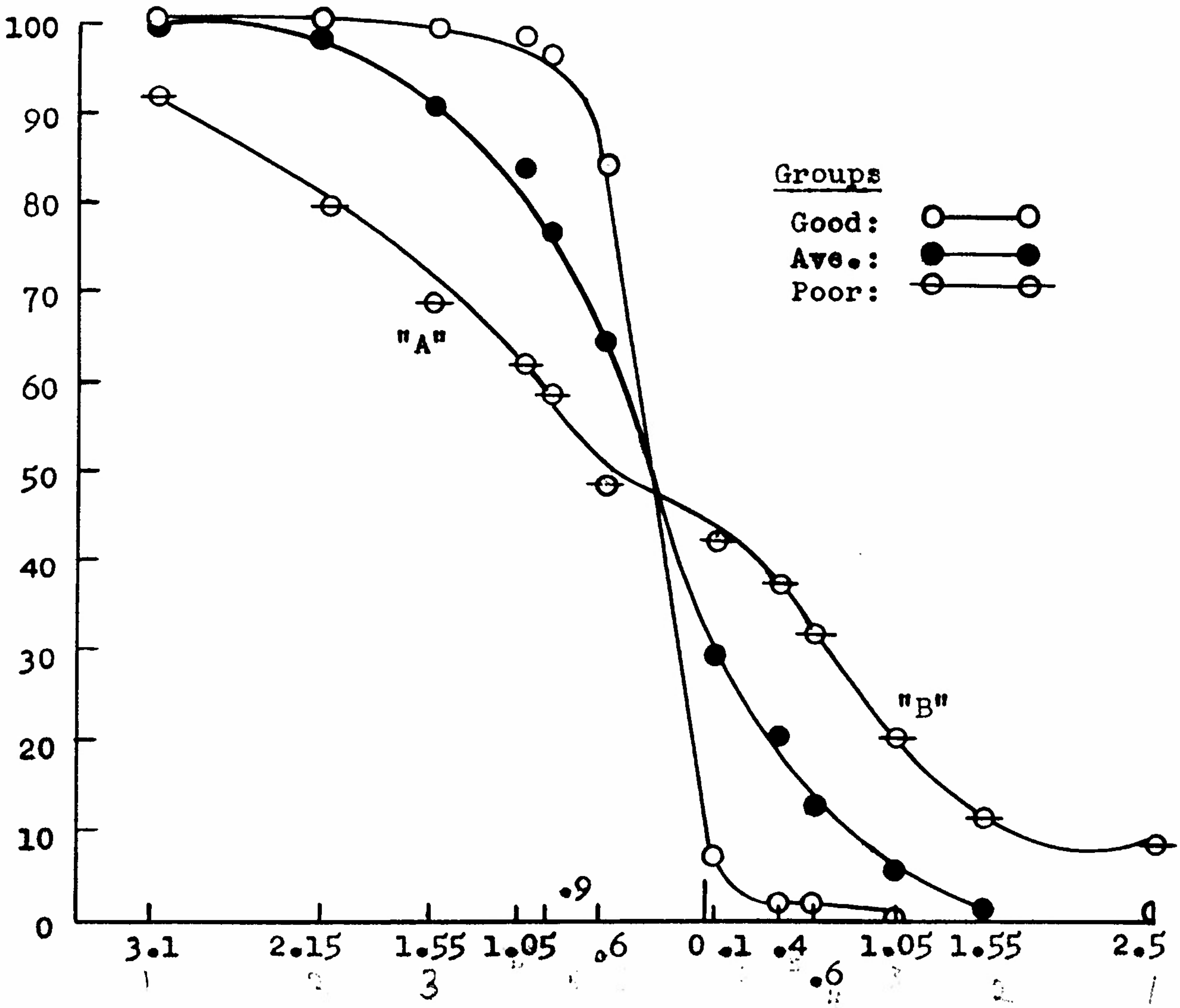
The explanation for the fact that the more difficult items (see last section of Table I) are not all highly discriminative is of course the fact that not only the "Poor" group, but the "Average" group also misses these items almost by chance. It may be pointed out that while the difference in per-cent correct is rather slight for a few items, these differences are nevertheless in the expected direction,--that is, the "Average" group excelled the "Poor" group on all 110 items.

PSYCHOPHYSICAL FUNCTIONS OF LOUDNESS:

If one relates percentage of correct responses against level of difficulty of items, it is seen that two complicating factors appear. Not only is there the expected difference between groups of differing ability, but the items which change in the "Louder" dimension are

inherently easier than those which change in the "Softer" dimension. Both of these factors will appear in Figure 7. Here the ordinate represents percentage of a group responding "Softer" for items at any level of difficulty. The baseline is laid off in decibels from zero difference (in the middle) toward each side: to the left are represented those items which changed in the "Softer" dimension by the number of decibels found on the baseline; to the right are represented those items which changed in the "Louder" dimension. The interpretation of any point on the three curves (let us take for example the point labelled "A") is thus as follows: for the "Poor" group, 68.2 percent of subjects (see ordinate) called "Softer" those items which changed 1.55 db in the "Softer" direction. The point labelled "B" indicates that 20% of the Poor group called "Softer" those items which changed 1.05 db in the "Louder" direction.

It is immediately apparent that the function in Figure 7 is of different form for the three groups; a somewhat closer inspection is necessary to observe that the data reveal a constant tendency on the part of all groups for "Louder" items to be easier than "Softer". In order to make this observation easier for the reader, the data of Figure 7 are re-graphed in Figure 8, A, B, C, with the "Louder" items rotated at 180°, so as to be juxtaposed with the "Softer" items. Figure 8-A clearly shows that an equivalent decibel difference for a "Soft" and a "Loud" item does not make the two items at all the same thing psychologically. On those items which were "Softer" by 1.55 db, for example, only 68.2% of the poor group responded correctly, but 88.3% of the same group responded correctly when the items were "Louder" by the same value of 1.55 db.



Decibel difference of items in the "Softer" direction

Decibel difference of items in the "Louder" direction

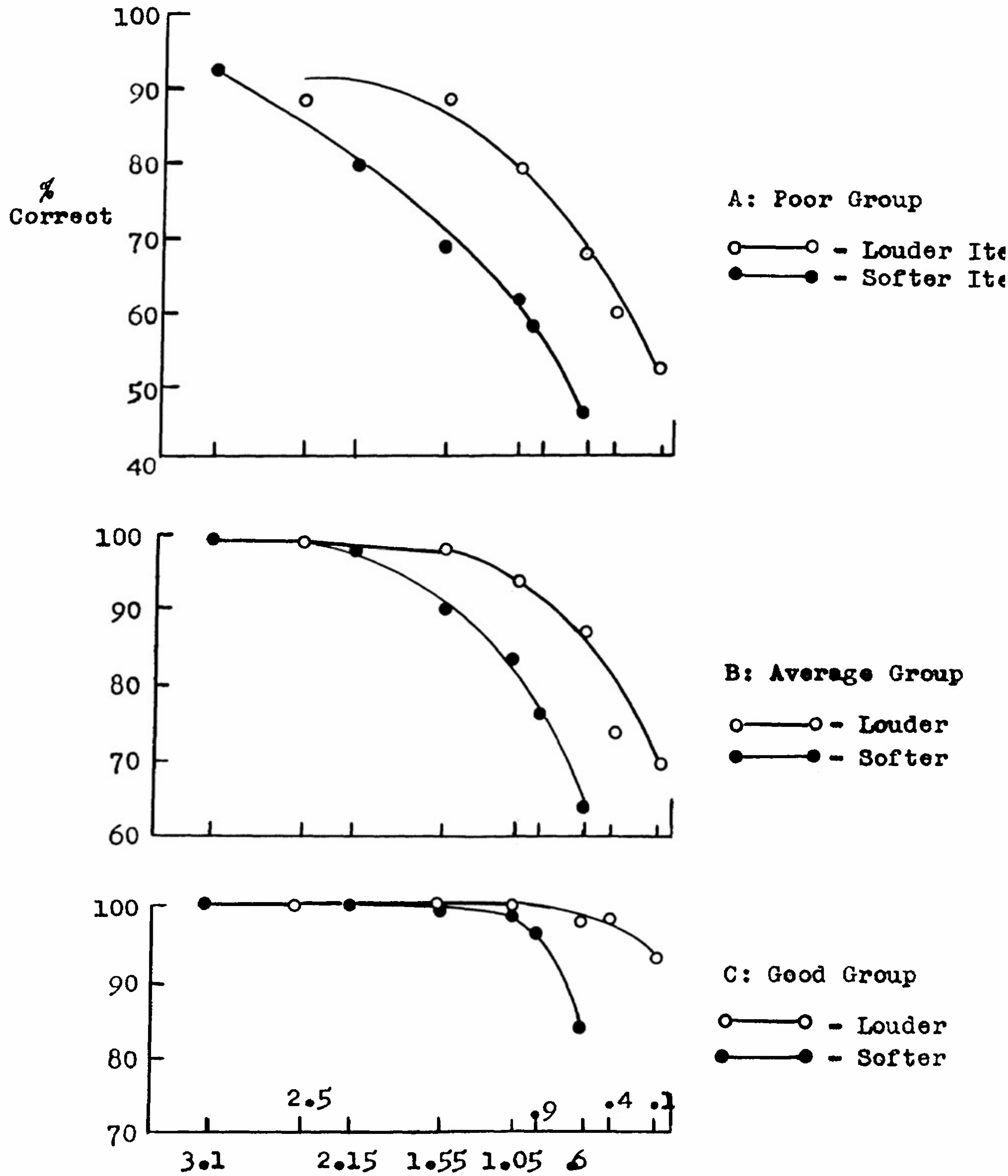
Ordinate: Percent of subjects calling item "Softer"

Abscissa: Items arranged by level of difficulty, ranging (left to right) from 3.1 db in the "Softer" direction to .6 db, and from .1 db in the "Louder" direction to 2.5 db difference.

Performance as a Function of Level of Difficulty

Fig. 7

Comparison of Performance on Softer
and on Louder Items



Abscissa: Level of Difficulty in Decibels

Fig. 8, A B C

A measure of this effect in terms of percent correct for a constant physical level of difficulty is found in the vertical distance between the curves; for any level except the easiest, the "Poor" group is roughly 15-20% better for the "Louder" items.

A more relevant measure, namely, that in terms of the change in decibel difference when percent of correct responses is held constant, is found in the horizontal distance between the curves. Reading across Figure 8-A at the 80% level, for example, we find that "Softer" items of 2.15 db difference, and "Louder" items of 1.05 difference, are psychologically equal; "Louder" items are throughout roughly as easy as "Softer" items in which the difference is sometimes a whole decibel wider.

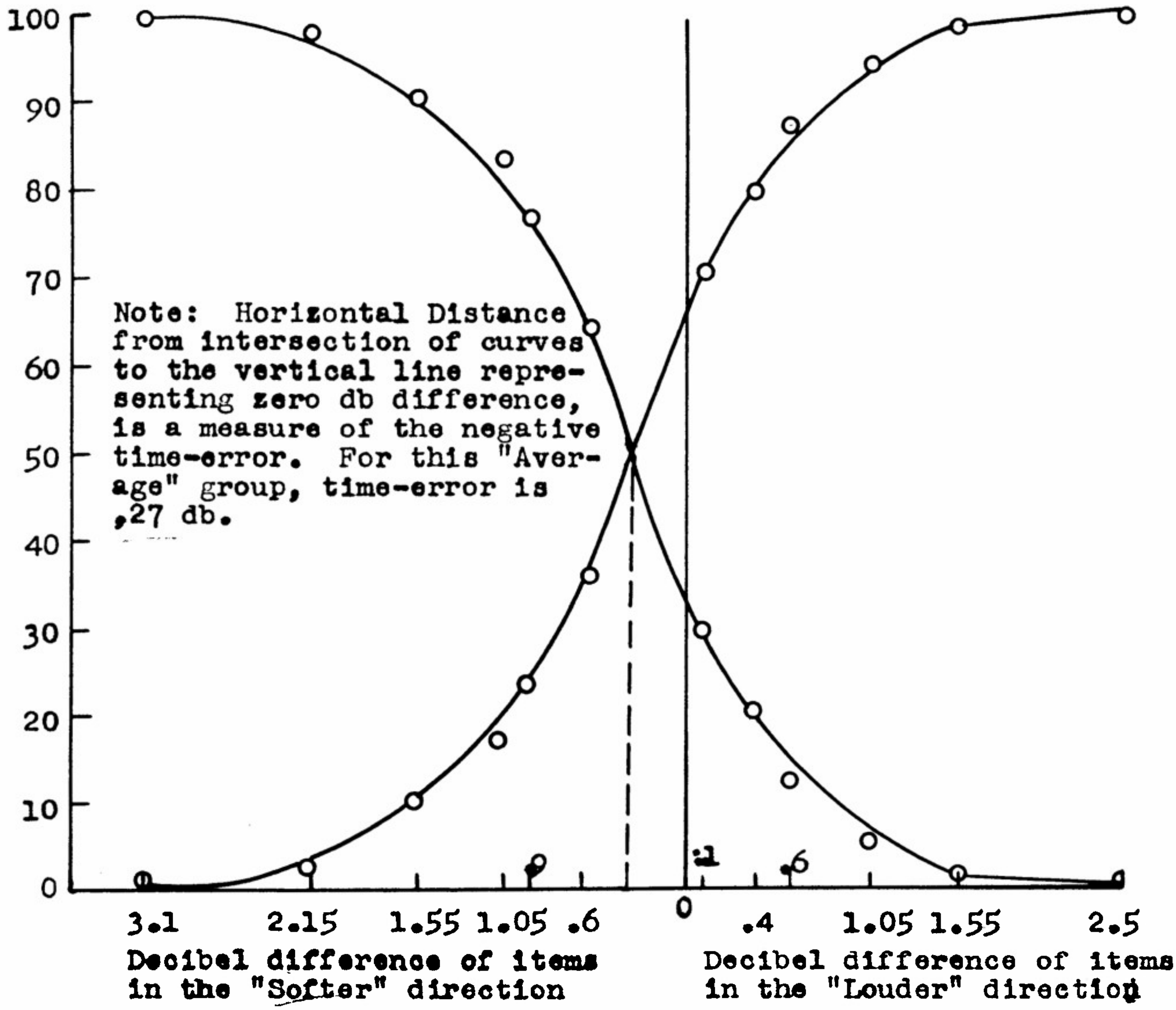
These effects are present likewise in both the "Average" and "Good" groups of Figure 8-B and C, though to a reduced extent. In the "Average" group, with constant level of difficulty (i.e., considering the vertical distance between curves), the "Louder" items are 10-15% easier, while for a constant psychological level of difficulty (horizontal distance between curves), the "Louder" items are easier by a quarter to a half decibel. The two curves in Figure 8-C can not be systematically interpreted due to the fact that only in the last few physical levels of difficulty does this "Good" group drop off appreciably from perfect performance. The two curves do, however, begin to diverge, the differences in both the vertical and horizontal directions becoming highly reliable.

THE TIME ERROR IN LOUDNESS DISCRIMINATION:

When the psychological method of Constant Stimuli is used for pure-tone loudness discrimination, it commonly appears that the second of the pair of tones seems louder, even though the two may be of the same intensity. This "Time-Error" as it is historically called was briefly considered in other terms in the foregoing graphs and discussion. We may here conveniently examine its precise amount.

Figure 9-A and B shows the data plotted as percent of responses against level of difficulty, just as in Figure 7, except that not only the correct responses to the "Softer" items, but also the correct responses to the "Louder" items are included. (Only the "Poor" and "Average" groups are included since the results of the "Good" group are almost identical with the data from the "Average".)

Consider first the pair of curves from the average data: The curve sloping down to the right is re-drawn from Figure 7, and shows the percent of "Softer" responses to both softer and louder items. The curve sloping up to the right indicates the percent of "Louder" responses to both softer and louder items. Of course for any group of items, the total percent of Softer and Louder judgments is 100. Now it would seem that, if no time-error were present, the curves should intersect where the 50% response level crosses the "zero db difference" vertical line. Yet this is obviously not the case. But if we drop a vertical to the baseline from the point where the curves do intersect (see broken line), we have a measure of the magnitude of the time error, in the horizontal distance from the point on the baseline to the zero db difference point. In terms of physical units, the magnitude is .27 decibel, in the "negative" direction (that is, the PSE, or Point of Subjective Equality) is less in magnitude than the Standard Stimulus.

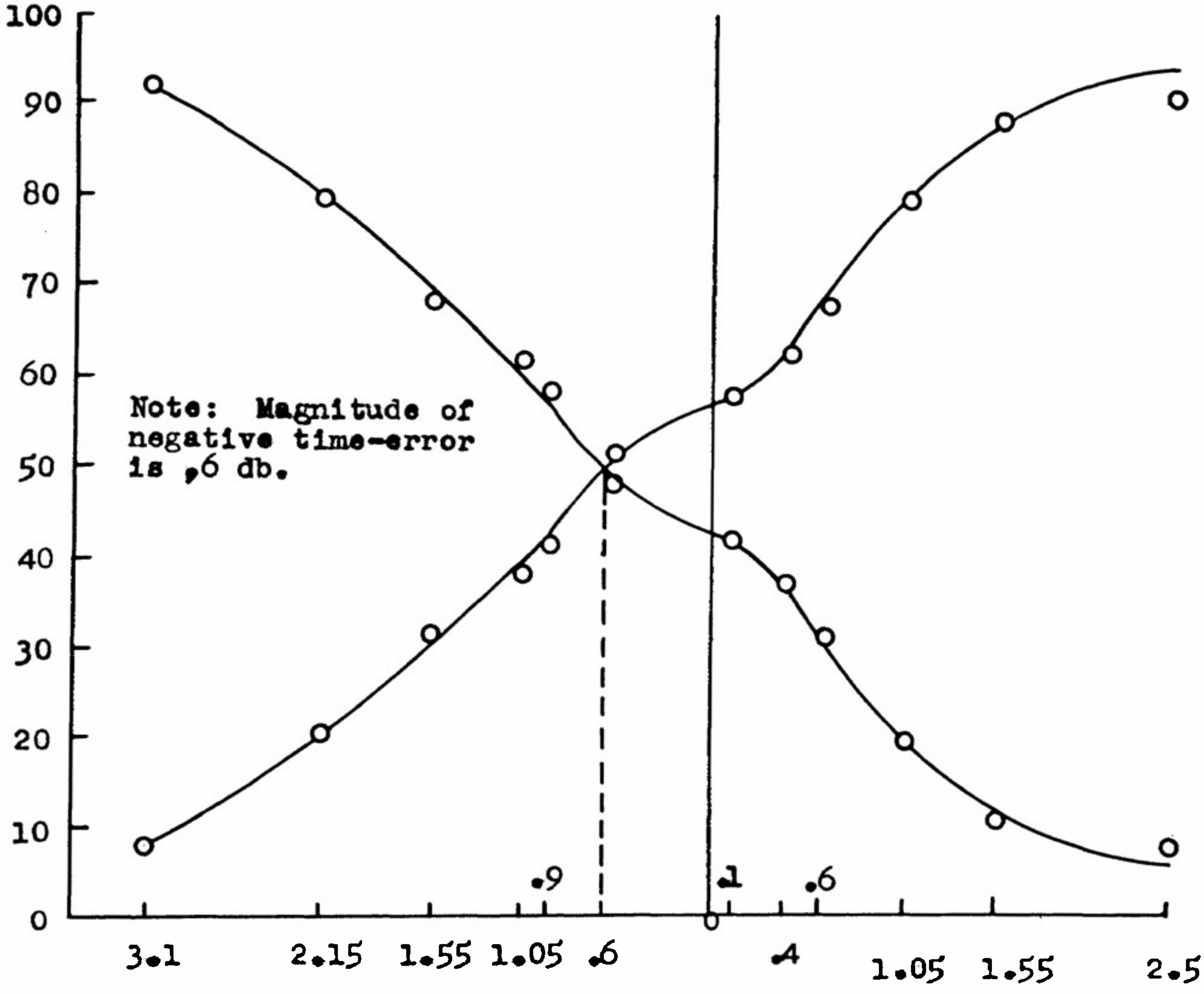


Abscissa: Items arranged by level of difficulty

Ordinate: Percent of subjects calling item "Softer", for curve sloping down to the right; for curve sloping up, ordinate represents percent of subjects calling item "Louder". The total percent for any level of difficulty is necessarily 100.

The Time-Error for the Average Group

Fig. 9A



Abscissa: Same as Fig. 9 A
Ordinate: Same as Fig. 9 A

The Time-Error for the "Poor" Group

Fig. 9 B

Figure 9-B is interpreted in exactly the same way. When we drop a vertical from the intersection of curves for the "poor" group, the magnitude of the negative time-error is .6 decibel.

We may briefly consider the nature of the time-error itself. It is not, for example, probable that it is due to some differential physiological factor, such as the explanation that the first tone produces auditory fatigue, or perhaps sensory adaptation, so that a second tone sounds louder in "successive comparison". Explanations in terms of the fading of an image, of "assimilation", and of "set" are likewise unsatisfactory for one reason or another. It is at least arguable that the explanation is a very simple one and can be put largely in neurological terms. The neural after-discharge of the first stimulus may continue for an appreciable interval, according to the well-known descriptions of Lorente de No of reverbatory neural "chains". Then by a relatively uncomplex process of neural summation, the neural effects of the second stimulus added to the still-continuing neural effects of the first, may combine to produce a higher central excitatory state (c.e.s.) than either the standard or variable stimulus could do alone. Neurologically speaking, the second stimulus <u>is</u> more intense than the first.

This explanation does no violence to, and indeed is generally supported by, what is known of the phenomenon as it is related to practice, interpolated material, and interval of time between standard and variable stimuli.

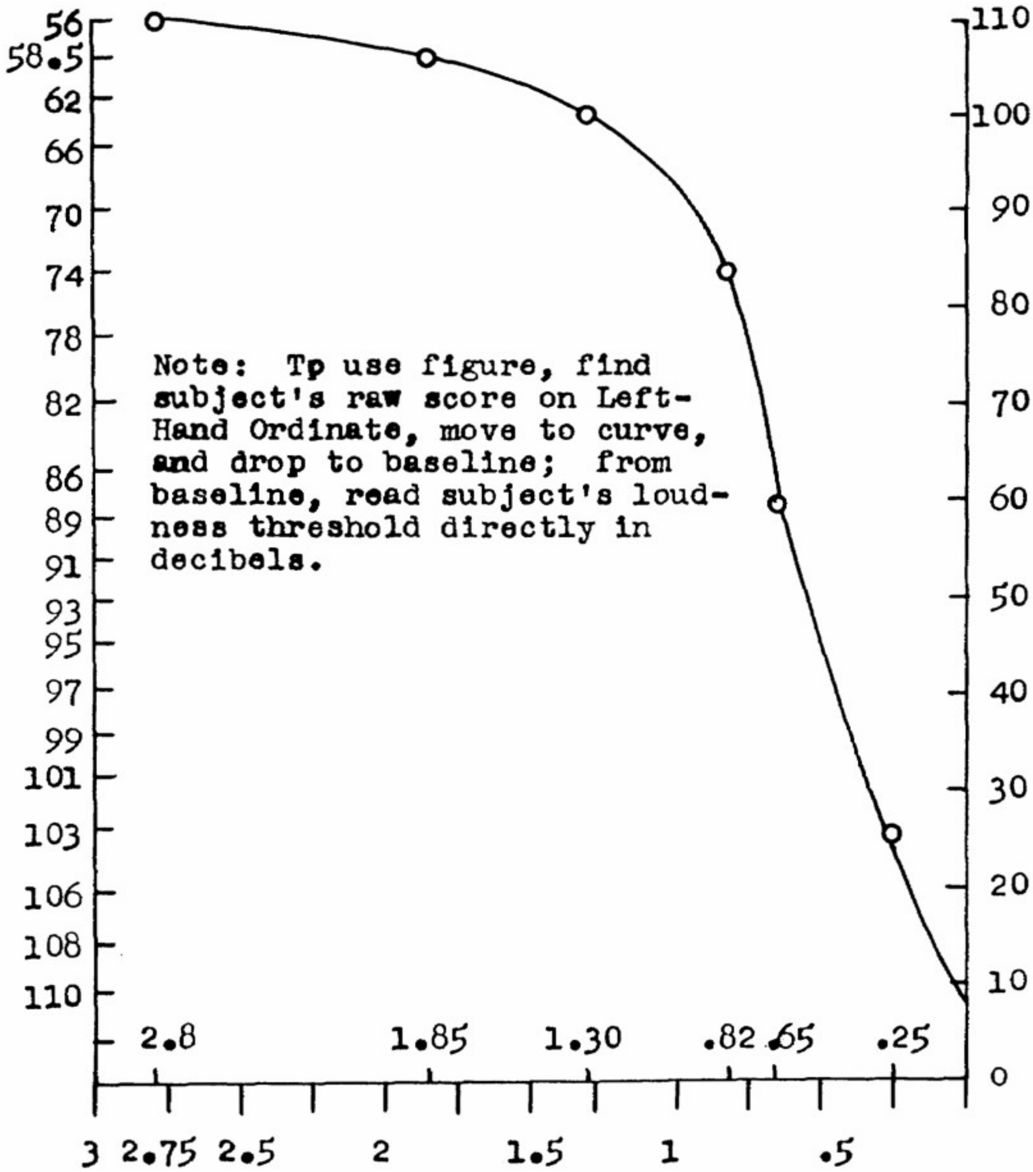
THE DETERMINATION OF DIFFERENTIAL LOUDNESS THRESHOLD FROM SCORE ON LOUDNESS TEST:

Figure 10 is a cumulative graph showing on the right hand ordinate the number of items at, or harder than, any particular level of difficulty ("Louder" and "Softer" items averaged). The graph is to be read in the following way: There are 110 items at or harder than the easiest level of difficulty (Ave. db difference = 1.85), and so on.

What we wish to derive from Figure 10 is some indication, from a subject's raw score, approximately what db difference he is able to discriminate. The theoretical argument for this derivation proceeds as follows:

There are 110 items. Now a raw score of 55, or chance correct, does not place the subject's threshold, since he could get 55 with no ability whatsoever. But if a subject is correct on, let us say, three-fourths of the items at the easiest level, then by a usual convention he ought to be given credit for that level. If a subject scores 55, we assume he got two of the four easiest items correct; but if he scores 56, we assume he got 3 of the 4 easiest items correct. This satisfied our criterion, gives him credit for the easiest level, and indicates that the number 56 should be entered on the Left Hand ordinate at such a point that where its horizontal extension intersects the curve, it does so at the point corresponding to the 2.8 level of difficulty on the baseline. This done, as the Left Hand ordinate of Figure 10 shows.

The point on the Left Hand ordinate which represents the raw score a subject must have before given credit for the second level of difficulty is found in a similar manner. If a subject scores 57, we assume he got all the easiest level correct. Then in order for him



Abscissa: Level of Difficulty (Loud and Soft items averaged)

Right Hand Ordinate: Number of items at <u>or</u> <u>harder than</u> each lvel of difficulty

Left Hand: Raw Score.

Nomograph for finding Loudness Threshold from Raw Score on Test

Fig. 10

to be given credit for the second level he must get three-fourths of that level correct. Of course he has gotten one-half of the second level correct by chance, so we merely add, to 57, one-fourth of the number of the items (6) in the second level, or 57 + 1.5 = 58.5. We therefore enter the number 58.5 on the Left Hand ordinate in the appropriate place. Any-one wishing to ascertain what a raw score of 58 means in terms of db threshold, then, has only to locate 58 on the left hand ordinate, move horizontally to the curve, and drop to the baseline, where it will be seen that a raw score of 58 is equivalent to a threshold of approximately 1.95 db.

So for credit at every level of difficulty, a subject must have marked all items at easier levels correctly, and have marked correctly three-fourths of the items at the level in question. In order for a subject to be given credit for the <u>hardest</u> level, he must have all 84 easier items correct plus three-fourths correct of the 26 hardest items, or 103.5. A score of greater than 103.5 by an extrapolation of the curve, gives credit for still finer discriminations, as the next paragraph explains.

On the assumption that all items at the same level of difficulty are actually of equal difficulty, we can lay off (in terms of raw scores) the distance on the Left Hand ordinate between any two adjacent levels of difficulty. This is done in Figure 10. Now if we wish to know what <u>any</u> raw score signifies in terms of how much db difference is discriminated (averaging "Louder" and "Softer" items), just as was explained for the second (1.85) level of difficulty, we have merely to find the raw score on the Left Hand ordinate, move horizontally to the curve, and at that point drop to the baseline, where the average threshold can be read directly in decibels.

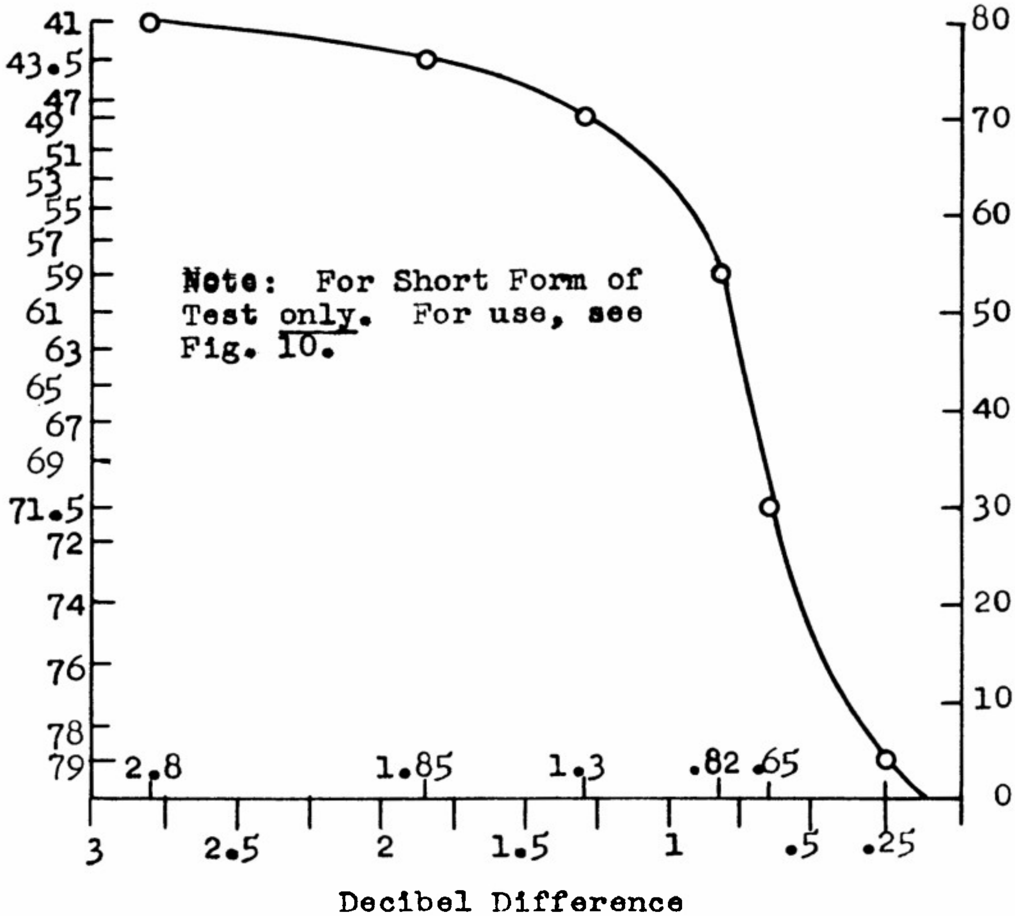
Several objections can be raised against this chain of reasoning, principally that "Louder" and "Softer" items should not be averaged, that items in physically similar groups are not necessarily psychologically equivalent, that a subject gets all of a certain level correct before he does better than chance on the next level, and that "chance" is too strictly interpreted. It is true that these considerations introduce an error of unknown amount; nevertheless, the procedure can be justified theoretically and is, moreover, about the only way to derive a threshold from a raw test score.

It is possible by the device of Figure 10 to say from the score a subject obtains on the loudness test, approximately how many decibels difference he can discriminate. Exactly the same reasoning underlies the construction of Figure 11. Here a subject's differential loudness threshold can be determined from his score on the short form of the test.

RELATION OF THE TEST TO SONAR PERFORMANCE:

No relationship between the test and the final grades of the New London Submarine Sound School could be discovered. There could not, however, be said unequivocally to be no relation to the more specific auditory performances involved, due to the nature of the criteria. For example, there were classes graduated in which almost no range existed on the final "Operating" score, while the "Final" score, which included all sub-marks, was so closely related to intelligence that any specific auditory factor could not be expected to appear.

It was decided that this activity would have to take the initiative in developing a number of tests in the New London Sound School, in cooperation with the instructors there, in



Abscissa: Same as Fig. 10
Ordinates: Same as Fig. 10

Nomograph for finding Loudness Threshold from Raw Score on Short Form of Test

Fig. 11

order to relate our preselection battery against a criteria toward with purely auditory factors would be more reasonably presumed to contribute.

A variety of tests were developed, standardized test of the ability to detect faint contacts at various levels of ownship noise, of the range through which target could be heard, of the ability to take accurate turn-counts under a variety of loudness levels and ratios of target-background noise, and the like. These standardized performance tests were constructed with the use of the Primary Listening Teacher (a training device which simulates a propeller in a noise background), the Advanced Listening Teacher (a device similar to the Primary Teacher but more carefully calibrated and capable of producing much more complex problems), and the JP sonic receiver-amplifier, using as a sound source either a launch moving in a regular pattern in the river, a moored launch, or in the laboratory, a Primary Listening Teacher coupled to the JP hydrophone input.

Not all of the specific performances mentioned above would be supposed to depend even in part upon good loudness discrimination; but some of them do seem to be so dependent. In one experiment, for example, a group of subjects was tested on the Advanced Listening Teacher, subjects being required to report "Contact" or "Lost Contact", as own-ship noise was lowered or raised whichever the experimental design called for. The reliability (test-retest) of the performance was quite satisfactory. For 28 ears, the Pearson product-moment correlation between score on this performance, and score on a routine presentation of our loudness discrimination test, was .40 $\pm$.01. All subjects were in the 3rd week of Sound School and had had individual and group practice on the instrument. The relation between contact

detection according to the Advanced Listening Teacher, and the loudness test is therefore by no means insignificant. Unfortunately, before these observations could be extended the instrument was moved to San Diego.

In another experiment, in which the Primary Listening Teacher was coupled as a sound source to the JP receiver-amplifier, a group of subjects (N = 19) who scored high on the loudness discrimination test, was compared with a group (N = 17) which scored poor on that test. Subjects were required to find the correct bearing when a problem was set up on the sound gear; the error in degrees was noted and averaged for each of 5 separate problems for each subject.

The average bearing error for the "Poor" group was 1.92 degrees; that for the "Good" group was 1.45 degrees. The difference of approximately half a degree is significant at the 5% level of confidence.

In another experiment using this same equipment but with different settings of the controls, a "Poor" group (N = 13) scored a mean of 1.72° error on 3 problems, a "Good" group (N = 34) scored a mean of 1.14°. The difference of .58° is reliable at the 99% level.[1]

From still another experiment with the use of this same equipment, a summary of the relation between loudness discrimination and

---

<u>Footnote 1</u>: A full report is in preparation on the experiments mentioned in this section as well as other experiments along the same line.

ability to report accurate bearings is contained in the following four-fold table:

	I	II
A:	1	17
B:	8	26

(Tetrachoric r = .51)

where A represents those subjects who reported accurate bearings (better than 1° error average), B those who did poorly (1° or more ave. error); I represents those subjects who failed (T-score lower than 40), II those who passed the loudness test. It is seen that of the 9 subjects who failed the loudness test, only 1 was able to report bearings accurate to 1°.

(It may be noted here that in a preliminary experiment a group of subjects who were blindfolded while performing the center-bearing experiment, made an average error no greater than that of a group who were free to use visual cues in splitting a bearing--in other words, it seems clear that not visual but auditory factors subserve the performance.)

On the basis of these and a body of similar results, this activity has been for some time recommending that those individuals who are far below average (more than 1 S.D. below the mean) on the loudness discrimination test, be considered for disqualification from submarine sonar training.

Within the last few weeks we have had access to scores from some training material which Dr. Adelbert Ford and his colleagues of NDRC have collected on students in the Submarine Sound School at San Diego. So far, these consist of phonograph records (reliability as yet unreported) of prop count, single ping, target classification, target differentiation, and

contact detection. Since all subjects in that school have been preselected in our laboratory we have been able to relate the latter three of these tests to our loudness test.

So far as the data is at present, the highest correlation obtained is .21, between our loudness test, v.s. contact detection. It remains to be seen whether in subsequent classes the relation is more pronounced.